Comprehensive functional analyses of expressed sequence tags in common wheat (triticum aestivum)

Manickavelu, A ; Kawaura, K ; Oishi, K ; Shin-I, T ; Kohara, Y ; Yahiaoui, N ; Keller, B ; Abe, R ; Suzuki, A ; Nagayama, T ; Yano, K ; Ogihara, Y

Abstract: About 1 million expressed sequence tag (EST) sequences comprising 125.3 Mb nucleotides were accreted from 51 cDNA libraries constructed from a variety of tissues and organs under a range of conditions, including abiotic stresses and pathogen challenges in common wheat (Triticum aestivum). Expressed sequence tags were assembled with stringent parameters after processing with inbuild scripts, resulting in 37,138 contigs and 215,199 singlets. In the assembled sequences, 10.6% presented no matches with existing sequences in public databases. Functional characterization of wheat unigenes by gene ontology annotation, mining transcription factors, full-length cDNA, and miRNA targeting sites were carried out. A bioinformatics strategy was developed to discover single-nucleotide polymorphisms (SNPs) within our large EST resource and reported the SNPs between and within (homoeologous) cultivars. Digital gene expression was performed to find the tissue-specific gene expression, and correspondence analysis was executed to identify common and specific gene expression by selecting four biotic stress-related libraries. The assembly and associated information cater a framework for future investigation in functional genomics.

DOI: https://doi.org/10.1093/dnares/dss001

DNA Research **19**, 165–177, (2012) doi:10.1093/dnares/dss001
Advance Access Publication on 14 February 2012

Comprehensive Functional Analyses of Expressed Sequence Tags in Common Wheat (*Triticum aestivum*)

Alagu Manickavelu[1,†], Kanako Kawaura[1], Kazuko Oishi[2], Tadasu Shin-I[2], Yuji Kohara[2], Nabila Yahiaoui[3], Beat Keller[3], Reina Abe[4], Ayako Suzuki[4], Taishi Nagayama[4], Kentaro Yano[4], and Yasunari Ogihara[1,*]

Kihara Institute for Biological Research, Yokohama City University, Maioka-cho 641-12, Yokohama 244-0813, Japan[1]; Genome Biology Laboratory, National Institute of Genetics, Mishima 411-8540, Japan[2]; Institute of Plant Biology, University of Zurich, Zollikerstrasse 107, 8008 Zurich, Switzerland[3] and Bioinformatics Laboratory, Faculty of Agriculture, Meiji University, 1-1-1 Higashi Mita, Kawasaki 214-8571, Japan[4]

*To whom correspondence should be addressed. Tel. +81 45-820-2405. Fax. +81 45-820-1901.
Email: yogihara@yokohama-cu.ac.jp

Edited by Satoshi Tabata
(Received 21 May 2011; accepted 30 December 2011)

Abstract

About 1 million expressed sequence tag (EST) sequences comprising 125.3 Mb nucleotides were accreted from 51 cDNA libraries constructed from a variety of tissues and organs under a range of conditions, including abiotic stresses and pathogen challenges in common wheat (*Triticum aestivum*). Expressed sequence tags were assembled with stringent parameters after processing with inbuild scripts, resulting in 37 138 contigs and 215 199 singlets. In the assembled sequences, 10.6% presented no matches with existing sequences in public databases. Functional characterization of wheat unigenes by gene ontology annotation, mining transcription factors, full-length cDNA, and miRNA targeting sites were carried out. A bioinformatics strategy was developed to discover single-nucleotide polymorphisms (SNPs) within our large EST resource and reported the SNPs between and within (homoeologous) cultivars. Digital gene expression was performed to find the tissue-specific gene expression, and correspondence analysis was executed to identify common and specific gene expression by selecting four biotic stress-related libraries. The assembly and associated information cater a framework for future investigation in functional genomics.

Key words: wheat; ESTs; annotation; transcription factors; miRNA; SNP; correspondence analysis

1. Introduction

Wheat provides 21% of food calories and 20% of protein to more than 4.5 billion people worldwide.[1] Demand for wheat in the developing world is projected to increase 60% by 2050. At the same time, climate change-induced temperature increases are estimated to reduce wheat production by 29%.[2]

The advent of new molecular genetic technology and the dramatic increase in plant gene sequence data have provided opportunities to underpin wheat breeding programmes in order to improve yield, grain quality, and disease resistance.[3] Many of these technologies have been designed to facilitate detection and understanding of the alterations in gene expression that accompany differential development or that result from the perception of changes to the environment. Expressed sequence tag (EST) projects provide a very useful and quick means of accessing gene sequence and expression information. When combined with breakthroughs in highly parallel designs for gene expression analysis, large-scale EST projects now offer new perspectives for understanding the molecular basis of important traits in plants of agricultural relevance.[4] EST sequencing projects have been completed or are under way for many plant species. These projects have provided useful tools for intragenomic[5] and intergenomic[6] comparisons, gene discovery,[7–9] molecular marker identification,[10] microarray

† Present address: ICARDA, PO Box 5466, Aleppo, Syria.

development,[11–15] and polyploid species genomic resource development.[16–18] As robot throughput increases and cost-per-read drops, determination of a sequence tag for a large proportion of genes is now reasonable using this random cDNA sequencing approach. For example, the availability of the complete genome sequence of *Arabidopsis thaliana* revealed that the 105 000 ESTs available at the end of 2000 were enough to tag 60% of the 25 500 genes.[19] The complete genome sequences of several plant species are known and the rate at which whole genomes are being sequenced is increasing. Correct annotation of these genomes remains problematic despite gene prediction algorithms becoming ever more sophisticated. While wheat genome sequencing is rapidly progressing (www.wheatgenome.org), a quicker and complementary approach to identifying a large number of wheat genes is EST and full-length cDNA sequencing. These resources will prove invaluable for annotating the genomes of wheat and other monocots and as substrates for transgenic improvement of crops. As in Arabidopsis and rice, these tools will prove to be critical in speeding up the genetic improvement of wheat.

DNA markers constructed from ESTs are effective since they are contained within an exon region of genes that are actually expressed. Examination of DNA sequence databases permits a direct search for sequence polymorphisms and thus molecular markers. These polymorphisms are typically single-nucleotide polymorphisms (SNPs) or small insertions–deletions (indels). More importantly, the SNPs are identified in EST sequences, thus the polymorphisms can be used to directly map functional, expressed genes, rather than DNA sequences derived from conventional RAPD and AFLP techniques, which are typically not genes. This has led to studies on linkage disequilibrium in genes to better characterize associations between phenotype and genotype.[20–22] To identify an SNP from an EST database, the database must be composed of ESTs derived from different genotypes, followed by alignment of the same EST sequences from different genotypes.[23] SNP markers rely upon the underlying redundancy within EST collections and assume that distinct genotype of a plant genome will be represented within a collection.

Earlier, we described extensive wheat EST resources including full-length sequences,[24–27] and the usage of wheat transcriptome analysis by making a custom microarray with ESTs.[28,29] Here, we extend our efforts and describe a collection of further ESTs and the complete functional analysis of the whole EST so far developed (~1 million ESTs). Our work is based on a set of cDNA libraries established from 51 different tissues of interest varied from growth stages and biotic and abiotic stresses in 10 different cultivars.

2. Materials and methods

2.1. Plant materials and cDNA library construction

Eighteen libraries from various growth stages, 25 libraries from abiotic stresses (cold, drought, saline, and mineral toxic), and eight libraries from biotic stresses (leaf rust, powdery mildew, and blast) were constructed from eight different wheat lines (Table 1). Out of 51 libraries, 20 libraries were newly constructed and included for this study. Double-stranded cDNAs were synthesized as previously described.[24] cDNAs were ligated with pBlueScript SK(+) digested with *Eco*RI and *Xho*I. After transformation by electroporation, transformed bacterial cells were initially cultured in the SOC medium for 1 h before culture at 37°C for 2 h in 2× LB medium. Cultured cells were stored in 20% glycerol at −80°C until use. Transformed bacteria were randomly selected and plasmid DNA was extracted.[24] Inserted cDNAs were sequenced from both ends using dye terminator cycle sequencing (Applied Biosystems, Foster City, CA, USA).

2.2. EST processing and assembly

The chromatogram files were base called and quality trimmed using PHRED[30] with default parameters. Vector, library linker-primer, and *Eco*RI adapter sequences were removed using CROSSMATCH. Repeat, ambiguous sequences (PHRED quality values <30) and poly (A) tails or poly (T) sequences (at most 10 bases) in the ESTs were trimmed. Subsequently, ESTs with sequences <30 bp were omitted from the final data set. The remaining high-quality sequence was used for further study. All sequence data are available from the DNA database of Japan (Table 1). The processed EST sequence files were combined and assembled into contigs using the CAP3 program[31] with a high and low stringency level (high 95% homology in a 20 bp overlap; low 80% homology in a 15 bp overlap). Default CAP3 settings include -p 90 -h 20; the custom parameter settings used were -p 85 -h 90. The CAP3 -p option specifies overlap per cent identity cut-off, while the -h option specifies the maximum alignment overhang percentage.

2.2.1. Sequence annotation Using the BLAST program (BLASTX with a search threshold of 1e−5), the sequences of the contigs were searched against seven databases (NCBI's nr; http://www.ncbi.nlm.nih.gov/genbank, Uniprot; http://www.uniprot.org, RAP-DB; http://rapdb.dna.affrc.go.jp, RGAP; http://rice.

Table 1. List and characteristics of cDNA libraries

Library name	Genotype	Stage	Condition	No. of EST	Accession number
whcs	CS	Callus	GS	11 505	CJ518205–CJ523460, CJ627048–CJ632007
whr	CS	Root	GS	19 227	BJ277129–BJ287630
whs	CS	Seedling	GS	13 356	HX000001–HX010004
whdl	CS	Seedling crown	GS	12 761	BJ221844–BJ231912
whh	CS	Spike at heading	GS	20 648	BJ255495–BJ266779
whf	CS	Spike at flowering	GS	21 106	BJ243195–BJ255494
whoh	CS	Pistil at heading	GS	20 736	BJ266780–BJ277128
whpc	CS	Anther at meiosis	GS	11 016	CJ576197–CJ580898, CJ682880–CJ687382
whhg	CSMT4B[a]	Anther at meiosis	GS	9669	CJ549536–CJ554132, CJ657247–CJ661657
whsh	CS	Young spikelet	GS	11 302	CJ730709–CJ736986
whyd	CS	Spikelet at late flowering	GS	14 708	BJ300204–BJ312233
whok	CS	DPA5	GS	12 159	CJ570869–CJ576196, CJ677747–CJ682879
whms	CSDT3DL[b]	DPA5	GS	12 036	CJ565425–CJ570868, CJ672462–CJ677746
whe	CS	DPA10	GS	19 200	BJ231913–BJ243194
whdp	CS	DPA20	GS	13 455	CJ523461–CJ529179, CJ632008–CJ637596
whsl	CS	DPA30	GS	15 522	BJ287631–BJ300203
whsp	CS	Seedling	GS	12 783	HX247045–HX247474
whca	CS(Sp5A)[c]	Seedling	GS	794	HX247475–HX257200
whkp	CSKmppd[d]	Seedling	Grown under continuous light	12 950	CJ554133–CJ559953, CJ661658–CJ667190
whkv	CS	Seedling	Grown under continuous light after cold treatment	15 360	CJ559954–CJ565424, CJ667191–CJ672461
whem	Kitakei1354	Dormant seed	With water supply	11 671	CJ539482–CJ544724, CJ647722–CJ652633
whei	Kitakei1354	Dormant seed	With water supply after wounded	11 743	CJ534326–CJ539481, CJ642661–CJ647721
whsc	Kitakei1354	Shoot	Cold treatment after excision of grain part	13 079	CJ586310–CJ591776, CJ692607–CJ697797
whsd	Kitakei1354	Shoot	Dehydration	11 897	CJ591777–CJ596845, CJ697798–CJ702615
whrd	Kitakei1354	Root	Dehydration	12 436	CJ580899–CJ586309, CJ687383–CJ692606
whv3	Valuevskaya	Shoot	3 days cold condition	10 069	CJ601934–CJ606680, CJ707296–CJ711731
whv	Valuevskaya	Shoot	16 days cold condition	11 087	CJ596846–CJ601933, CJ702616–CJ707295
whva	Valuevskaya	Shoot	ABA treatment	10 631	CJ606681–CJ611586, CJ711732–CJ715867
whvd	Valuevskaya	Shoot	Five days dehydration	11 767	CJ611587–CJ616926, CJ715868–CJ720922
whvh	Valuevskaya	Shoot	Heat shock treatment	10 090	CJ616927–CJ621531, CJ720923–CJ725419
whvs	Valuevskaya	Liquid cultured cells	Liquid cultured cells	12 327	CJ621532–CJ627047, CJ725420–CJ730708

Continued

Table 1. Continued

Library name	Genotype	Stage	Condition	No. of EST	Accession number
whrs6[e]	CS	Root	Salt stress for 6 h	13 110	HX010005–HX019847
whss6[e]	CS	Leaf	Salt stress for 6 h	13 312	HX019848–HX030180
whrs24[e]	CS	Root	Salt stress for 24 h	12 949	HX030181–HX040252
whss24[e]	CS	Leaf	Salt stress for 24 h	12 487	HX040253–HX050054
whatl[e]	Atlas66	Root	No treatment	20 519	CJ773323–CJ797201
whatlal[e]	Atlas66	Root	50 mM Al for 6 h	28 795	CJ822818–CJ848636
whsct[e]	Scout66	Root	No treatment	27 717	CJ797202–CJ822817
whsctal[e]	Scout66	Root	50 mM Al for 6 h	29 824	CJ848637–CJ872807
whhb[e]	Halberd	Root	No treatment	22 488	HX124648–HX143755
whhbb[e]	Halberd	Root	10mM boric acid for 24 h	22 522	HX143756–HX163093
whcr[e]	Cranbrook	Root	No treatment	22 680	HX163094–HX182808
whcrb[e]	Cranbrook	Root	10mM boric acid for 24 h	22 616	HX182809–HX201765
whthls[e]	Thatcher	Seedling	Infected with leaf rust	30 307	CJ872808–CJ896490
whthkles[e]	NILThatcher	Seedling	Infected with leaf rust	24 701	CJ896491–CJ919993
whchan[e]	Chancellor	Seedling	Infected with powdry mildew	29 281	CJ919994–CJ944155
whchu[e]	NILChancellor	Seedling	Infected with powdry mildew	28 799	CJ944156–CJ968175
whnr[e]	Norin4	Seedling	No treatment	34 415	HX050055–HX071918
whnrpr48[e]	Norin4	Seedling	Infected with blast strain Pr48 at 23°C for 4 days	23 987	HX071919–HX084716
whnrpr58r[e]	Norin4	Seedling	Infected with blast strain Pr58 at 23°C for 4 days	36 360	HX084717–HX106894
whnrpr58s[e]	Norin4	Seedling	Infected with blast strain Pr58 at 27°C for 4 days	30 797	HX106895–HX124647
			Total	894 756	

CS, Chinese Spring; GS, growth stage; DPA, days to post-anthesis; NIL, near-isogenic lines.
[a]Mono-telosomic 4BS of CS.
[b]Ditelosomic 4BS of CS.
[c]Spelta5A chromosome substituted in CS.
[d]Near-isogenic line.
[e]Newly constructed libraries.

plantbiology.msu.edu, Tair9; http://www.arabidopsis.org, MaizeGDB; http://www.maizegdb.org, and Brachypodium database; http://db.brachypodium.org). According to Ewing *et al.*,[7] only contigs were taken for further analyses. The gene ontology (GO) terms[32] of each contig was derived by InterProScan.[33] The GO terms were then converted into GO slim term using EBI website (http://www.ebi.ac.uk/QuickGO/) by written perl-script for this purpose. Open reading frames (ORFs) were searched by translating sequences into amino acids by six frames (three per frame in the plus and minus strands).

2.3. *Transcription factor*

The PlnTFDB[34] containing 29 473 sequences of plant genes involved in transcriptional control was used to mine our data by local BLAST. The default parameters mentioned in the database was used for prediction (Filter 'on', gapped alignment 'on', substitution matrix 'blosum 62', *E*-value $\leq 1e-10$). We included two meta-rules in our classification scheme: (i) if a protein harbours domains characteristic of a transcription factor (TF) family and a transcriptional regulators (TR) family, we assigned it to the TF family, (ii) when the protein of interest contains domains characteristic of more than one TF family or more than one TR family, it was assigned to the family to which its characteristic domains matched with the lowest *E*-value.

2.4. *Full-length cDNA*

The contigs were classified as full length if it aligned with our full-length cDNA data.[27] BLASTN

searches of the contigs against the full-length sequences yielded a candidate hit list (*E*-value $\leq$1e−100) of putative full-length sequences that either covered the start and stop codon of the subject sequence or possessed sufficient sequence up/down-stream of the match to contain putative start and stop signals. In a few instances, some contigs covered all but the start methionine, and were also included as full-length sequences. In addition, contigs were aligned (*E*-value $<$1e−5, $\geq$98% query coverage, $\geq$98% identity) with barley full-length sequences[35] to know the similarity as well as the full-length nature.

2.5. *miRNA analysis*

To identify conserved miRNA in wheat, contigs were annotated with the plant small RNA regulator target analysis database (http://plantgrn.noble.org/psRNATarget/) containing small RNA of 15 plant species including wheat.[36] This database contained 2192 published miRNA sequences, including 32 from *Triticum aestivum*, 148 from sorghum, 496 from *Oryza sativa*, 319 from maize, and 224 from *A. thaliana*. Potential targets were predicted according to the rules applied by[37,38]: (i) the number of allowed mismatches at complementary sites between miRNA sequences and potential mRNA targets is four or fewer; and (ii) no gaps are allowed at the complementary sites.

2.6. *SNP discovery*

Sequence variants or SNPs were mined in wheat contigs with two criteria, and perl-scripts were written for each category. In the first criterion, only contigs with $\geq$4 ESTs were selected, and SNPs were declared only when there was no mismatch, no gaps, or N's were admitted before and after an SNP site; in addition, the alternative base to the consensus sequence was present at least more than twice in an alignment. In the second criterion, the SNPs were mined only in the significant sequence of contigs that was worked out by counting the nucleotides of either end of the contigs containing a minimum of four EST members. To find the SNP between cultivars, in addition to the above parameters, the contigs containing the minimum of two consistent EST from the same cultivar were selected. For homoeologous SNPs, the contigs containing the minimum of four EST from the same cultivar were chosen. The visual inspection of SNP was carried out using Tablet software.[39]

2.7. *Digital gene expression and correspondence analysis*

For statistical analysis of gene expression profiles, contigs harbouring five or more constituents were selected from 37 138 contigs. Similarities between contigs or libraries were estimated using Pearson's correlation coefficient.[40] Hierarchical clustering was applied to compare EST expression profiles among the 51 wheat tissue/treatments and libraries. Expression profiles are displayed based on the number of constituents in a contig (from 0 to 4647; red intensity), along with an increasing number of constituents. Contigs specific to DREB (dehydration-responsive element binding), NAC (nitrogen assimilation control), OMT (*O*-methyl transferase), and miRNA 172 were selected to show the differential gene expression in cultivars and growth stage.

Correspondence analysis (CA) was carried out by selecting four disease-related libraries (*whthls*, *whthkles*, *whchan*, and *whchul*; Table 1) as per the procedure detailed in Hamada *et al*.[41] and visualized by a custom-build viewer (available based on request).

3. Results

3.1. *cDNA library construction and assembly*

Our previous studies carried out the construction and analysis of 31 libraries.[26] Here, we have reported the further addition of 20 libraries and their combined analyses for comprehensive view of wheat EST. At the maximum, we have accumulated $\sim$1 million ESTs (Table 1). The libraries were generated from developmental stages and stresses. After trimming low-quality bases, vector sequences, and shortness ($<$30 bp), 0.68 million ESTs were used for CAP3 assembly under stringent conditions resulting in 37 138 contigs and 215 199 singlets. When assembled using relaxed settings in CAP3, 65 426 contigs and 66 875 singlets were obtained. The high stringent condition was chosen to achieve a more complete isolation of individual paralogues, orthologues, and homoeologues compared with using a low stringent condition.

The total sequence of transcript assemblies in stringent parameter settings, containing both the singlets and contigs, developed in this study was 125.3 Mb with the GC% of 51.9%. This is the maximum transcriptome sequences developed in any plant species. The GC% value is similar to rice but less than that reported in the wheat 3B chromosome exon coding sequence.[42] The length of the singlets varied from 31 to 884 bp with an average of 430 bp. The maximum singlets were grouped under 500−600 bp lengths (Fig. 1). As we found a large number of singlets, we subsequently discriminated the singlet contribution among the 10 cultivars (Fig. 1B). There was no correlation observed between the number of ESTs and the singlets. However, the stress-related libraries from four cultivars contributed to 55% of the total singlets. The contig

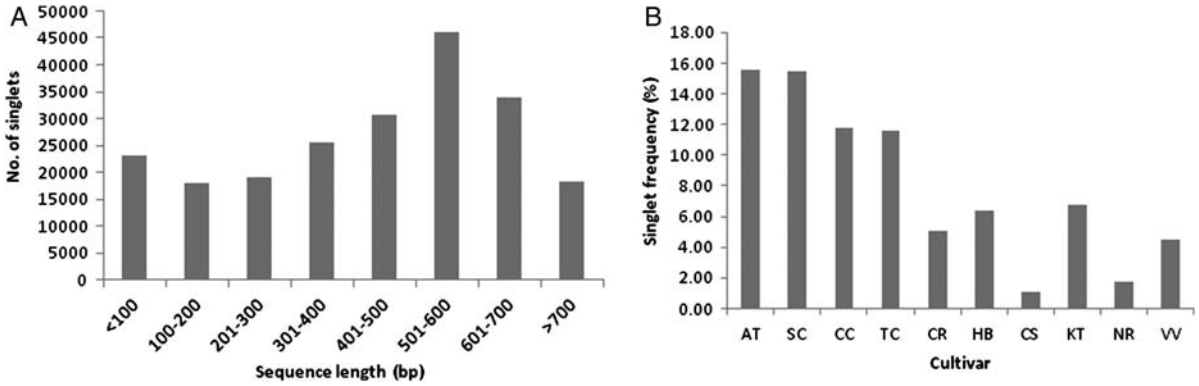

Figure 1. Analysis of singlet sequence length and their genotype-wise distribution. (A) Sequence length distribution of singlet. (B) Genotype-wise frequency (%) of singlet (AT, Atlas; SC, Scout; CC, Chancellor; TC, Thatcher; CR, Cranbrook; HB, Halberd; CS, Chinese Spring; KT, Kitakei1354; NR, Norin4; VV, Valuevskaya).

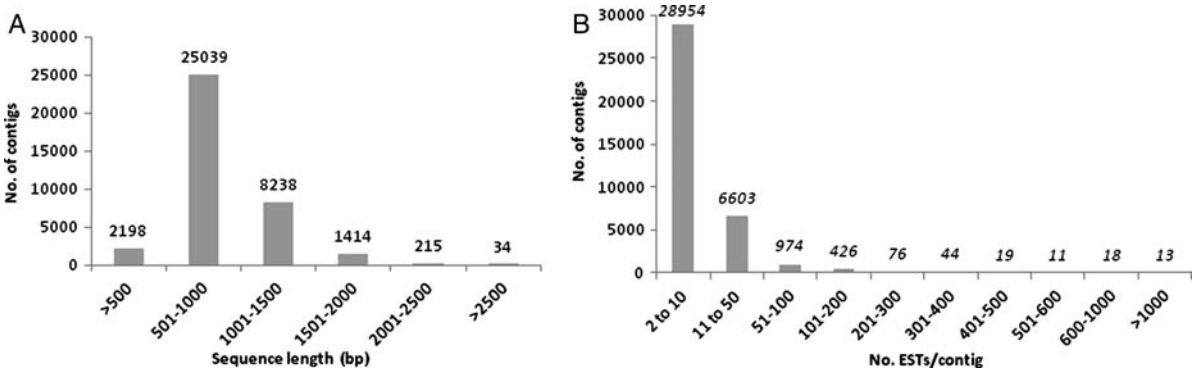

Figure 2. Distribution of contig length and their EST member constitution. (A) Sequence length frequency of contigs. (B) Number of EST members in contigs.

length ranged from 46 to 3960 bp with an average of 879 bp, and ~70% of the contigs extended from 501 to 1000 bp (Fig. 2A). The number of ESTs grouped in each contig varied between 2 and 4647, with 78% of the contigs containing 2–10 EST members (Fig. 2B).

3.2. *Functional annotation*

The contig resulted from stringent parameter assembly was used for functional annotation. The function of each contig was derived after annotation with rice, Arabidopsis, maize, and Brachypodium databases, in addition to the protein sequences in the GenBank nr database (BLASTX; *E*-value $<1e-5$). The recently sequenced Brachypodium was included for its close originated relation with wheat. On annotation, maximum similarity was observed in Brachypodium followed by rice (Table 2). In rice, further annotation was carried out to find the chromosome-wise sequence similarity and identified that chromosome 1 is having much co-linearity followed by chromosome 3 (Fig. 3). On overall annotation, ~3500 genes were found to have no similarity, suggesting new genes in our data. To further validate these new genes, updated tentative consensus sequences from the DFCI wheat gene index (http://compbio.dfci.harvard.edu/cgi-bin/tgi/gimain.pl?gudb=wheat) were annotated and resulted in the same number of new genes (~3500), confirming the importance of our new EST assembly and analysis in wheat. To estimate the total number of full-length cDNAs in our collection, we searched our contig data against our 11 902 full-length cDNA data. With the stringent criteria of >95% similarity and the expected cut-off value of $<1e-100$, we found ~7000 contigs that were full length in nature, further validated by identification of high similarity of wheat contigs with barley full-length sequences, indicating the robustness of our data and their applications to wheat functional genomics.

We also analysed length and ORF distribution in the contigs for both plus and minus strands. To obtain meaningful results, only contigs with >5 ESTs were selected and ORFs were identified. GO annotation of the wheat contigs was performed on the basis of ORF mining of the data. The GO terms were organized into three categories representing molecular functions, biological processes, and cellular

Table 2. Annotation of wheat contigs

Database and Species	URL	No. and percentage of similarity
RAP-DB (build 5) (Rice)	http://rapdb.dna.affrc.go.jp	32.405 (87.26%)
RGAP (Rice)	http://rice.plantbiology.msu.edu	32.430 (87.32%)
TAIR9 (Arabidopsis)	http://www.arabidopsis.org	30.504 (82.14%)
MaizeGDB	http://www.maizegdb.org/	31.206 (84.03%)
Brachypodium Database	http://db.brachypodium.org	32.522 (87.57%)
All databases		33.909 (91.31%)
Total contigs		37.138

The updated (until May 2011) sequence was retrieved and similarity search was carried out.

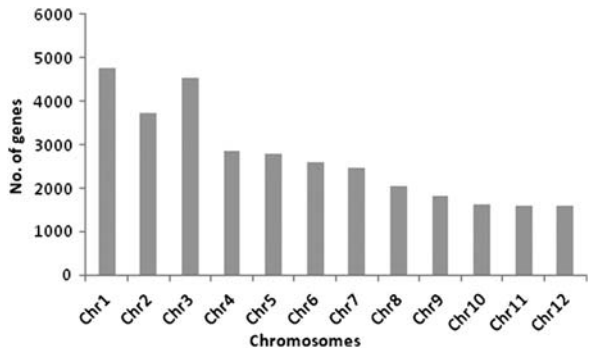

Figure 3. Sequence similarity of wheat contigs with rice genome. Based on the result, the contig was grouped in rice chromosome wise.

components.[32] The sum of the wheat contigs per category did not add up to 100% as some contigs were classified into more than one category. Of the total contig set, 21 125 (56%) were annotated into the molecular function category (describing the biochemical activity performed by the gene product), 13 354 (36%) into the biological process GO category (describing the ordered assembly of more than one molecular function), and 13 356 (36%) into the cellular component GO category (describing subcellular compartments of a cell) (Fig. 4). Among the molecular function, the most highly represented categories were binding, catalytic activity, redox activity, and structural activity (Fig. 4A). Among the biological processes, the largest proportion of functionally assigned contigs fell into metabolic, transport, and translation processes, while redox activity, biosynthetic process, regulation, phosphorylation, and transcription comprised 34% of the contigs (Fig. 4B). For the cell component category, almost all contig sequences were annotated into the cell–cell subcategory, 28% into the membrane category, and 23% into the intracellular category (Fig. 4C). Together, all three GO categories accounted for ~82% of the assigned wheat contig set.

The role and importance of TF lead to mining of our data and resulted in 1183 contigs containing either single or multiple transcription factors. Among the TFs, the *CCAAT* family was found in as many as 69 contigs. miRNA target sequence analysis of the wheat transcriptome identified different miRNA target sequences in 5180 contigs which ranged from 19 to 24 nt long. The majority of the small RNAs are 20–24 nt long, which is a typical range for dicer-derived products; the 21-nt class is predominant. Among species-specific miRNA, rice had maximum homology followed by maize and *Medicago truncatula* (Fig. 5). The number of hits for each species is roughly proportional to the number of sequences for that species in the database. Due to the limited number of wheat miRNA sequences in the database, there was only 200 contigs with wheat-specific miRNA target sequences. Among miRNAs, miRNA 395, 172, and 164 target sequences alone were found in 831 contigs, showing the relative abundance of these miRNA target sequences in wheat.

3.3. *Sequence polymorphism/SNP mining*

The SNPs were mined in our large-scale wheat transcriptome data by applying both relax and stringent criteria. For both criteria, to find reliable SNPs, the common conditions of SNP should be present at a given position when there is no mismatch present 2 bp before or after the SNP site. A total of 51 067 SNPs were detected from 20 609 contigs using the first criterion, resulting in the identification of an SNP site every 96 bp. This value is considerably higher than those reported for earlier studies of wheat.[23,43] Hence, the second criterion was applied with an interest to differentiate the homoeologous SNPs from intergenome SNP by calculating the significant sequence length of each contig before mining the SNPs. This approach avoided finding SNPs on either end of the contigs and resulted in the identification of only 6352 SNPs in the wheat contigs. Further classification of the SNPs present between cultivars found 3515 SNPs with a frequency of one SNP per 614 bp. As there were genome constituents of wheat (three homologous genomes) and selective gene expression among the three genomes, we examined the SNP within each cultivar and found 2837 SNPs with an SNP site every 470 bp. The overall SNP frequency based on the stringent criteria was one SNP per 483 bp. Transitions (70%) were more frequent than transversions (30%). As expected, a significant positive correlation ($P < 0.05$) was

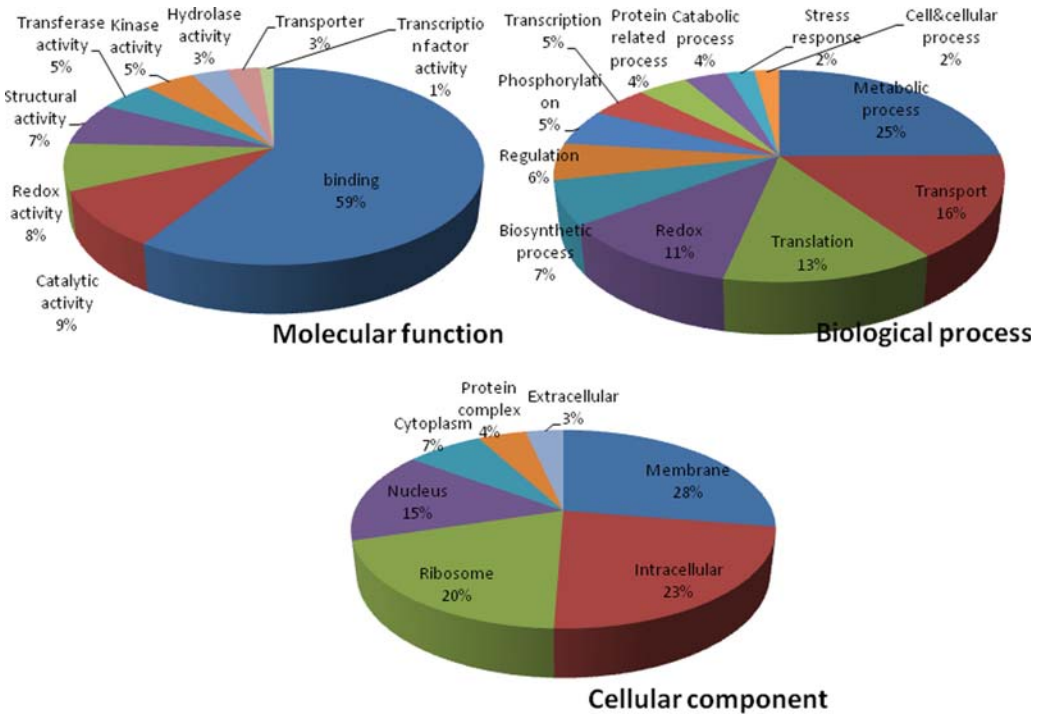

Figure 4. Functional classification of contig sequences based on GO categorization. Sequences were evaluated for their predicted involvement in molecular function, biological process, and cellular component.

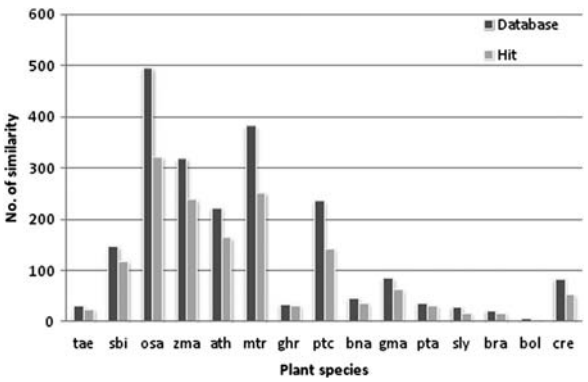

Figure 5. *miRNA* target sequence analysis in wheat contigs. The database bars indicate the available *miRNA* in the database and the hit bars indicate the number of wheat genes having *miRNA* target sequence. tae, *Triticum*; sbi, *Sorghum bicolor*; osa, *Oryza sativa*; zma, *Zea mays*; ath, *Arabidopsis thaliana*; mtr, *Medicago truncatula*; ghr, *Gossibium hirsutum*; ptc, *Populus trichocorpa*; bna, *Brassica napus*; gma, *Glycine max*; pta, *Pinus taeda*; sly, *Solanum lycopersicum*; bra, *Brassica rapa*; bol, *Brassica oleraceae*; cre, *Chlamydomonas reinhardtii*.

observed between the number of SNPs detected in a contig and the number of reads present in that contig. The cultivars, except Chinese spring and Norin4, contain more inter-cultivar sequence variation than homoeologous SNPs. Among cultivars, Halberd, Valuevskaya, and Cranbrook contain almost the same number of SNPs in both cases. Cultivars Atlas, Kitakei1354, Scout, and Thatcher had much less homoeologous sequence variation than cultivar differences (Supplementary Fig. S1).

3.4. *Digital gene expression*

EST frequencies approximate the message abundance in the mRNA population used to construct a cDNA library. We have already attempted to make tissue expression maps of a large number of ESTs from stress-related libraries for *in silico* screening of stress responsive genes in wheat.[26] Here, we aimed to determine the global gene expression of wheat from 51 cDNA libraries, including growth stages and biotic and abiotic stresses. Contigs containing >5 ESTs were subjected to a correlated clustering analysis[7] to compare expression profiles in the different libraries. When the result was displayed in the form of a dendrogram (Fig. 6), many libraries with similar origins agglomerated together. All four libraries derived from root tissues treated with boric acid and aluminium united. In the same manner, biotic stress (leaf rust, powdery mildew, and blast)-related tissues were grouped in the same clade. The tissues collected from cv. Valucvskaya, which mainly undergo abiotic stress, were clustered separately.

To further determine the tissue-specific gene expression of select genes, digital gene expression was carried out for DREB and NAC TFs, OMT gene,

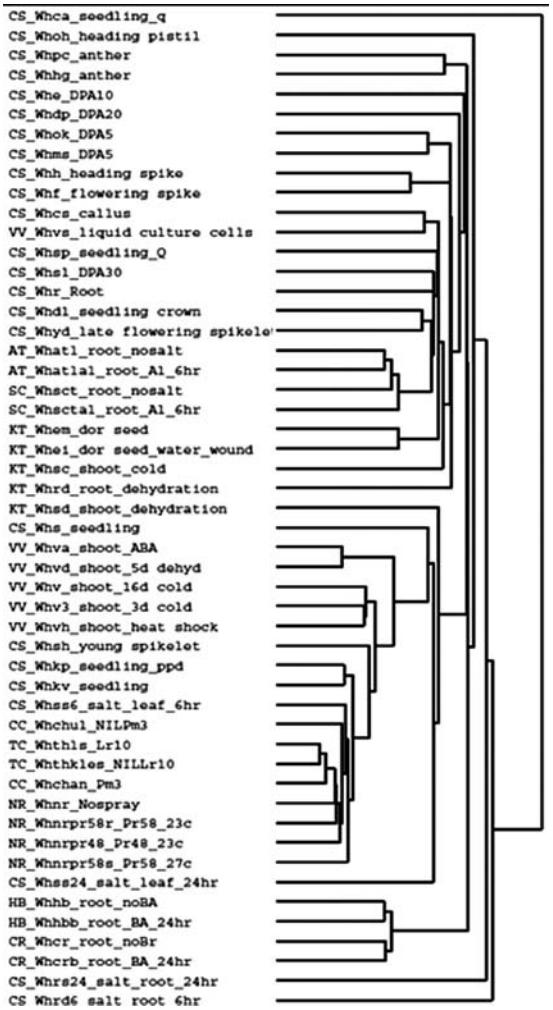

Figure 6. Correlated clustering of wheat cDNA libraries based on gene expression (AT, Atlas; SC, Scout; CC, Chancellor; TC, Thatcher; CR, Cranbrook; HB, Halberd; CS, Chinese Spring; KT, Kitakei1354; NR, Norin4; VV, Valuevskaya).

and miRNA 172 targeting site genes in wheat transcripts (Supplementary Fig. S2). The DREB genes are expressed mainly in dehydration-related tissue libraries and the spatial expression was mostly at root tissues. In the case of NAC TF genes, as expected, the expression was only noted in salt-treated libraries and the expression level was greater in root tissue followed by shoot, spikelet, and seed. Interestingly, some biotic stress-related libraries also expressed NAC TF. Based on the preliminary result obtained from our previous study on OMT against self-defence in wheat,[29] the OMT-related contigs were mined and its gene expression was analysed. We found ubiquitous expression of OMT genes irrespective of stress, suggesting its defence role against stress. While mining of miRNA in wheat transcriptome, we found an abundance of miRNA 172 target sites. The digital gene expression analysis showed its abundance in all tissues and under all treatments in wheat.

In addition to digital expression, a new method of gene discovery and/or gene expression based on CA was carried out to determine specific and common gene expression between libraries or treatments. To identify the common molecular plant–athogen interactions, four libraries constructed for leaf rust and powdery mildew diseases were selected and examined by CA. Contigs with more than or equal to four ESTs were selected, to identify specific genes for leaf rust and powdery mildew diseases, in addition to common disease resistance- and susceptibility-related genes (Supplementary Table S1). The number of genes expressed for powdery mildew outplayed the leaf rust disease. However, the number of disease susceptibility-related genes was more in leaf rust, suggesting different disease reaction mechanisms among diseases in wheat. Overall around 100 new genes were identified from these four disease-related libraries, which could have immense value for future research of molecular plant–pathogen interactions.

4. Discussion

Global wheat transcriptome analysis was carried out by accumulating ~1 million ESTs from 51 cDNA libraries. In comparison with other studies of wheat which were biased towards one stress or growth stage of a few cultivars,[3,26] here we used all growth stages, and biotic and abiotic stresses for 10 different cultivars. The work flow of EST assembly and further analyses were summarized in Fig. 7. Many perl-script programs were written specifically for processing the ESTs, resulting in a 24% reduction in total ESTs. The stringent parameter in EST assembly resulted in more singlets compared with contigs which helped for further analysis, i.e. SNP mining.[23] The average length of the contigs (879 bp) is higher than other studies,[3,23,44] and ~80% of the contigs containing 2–10 EST members had homoeologous or paralogous genes. To determine the EST member contribution to the contig, we further classified the data into stress- and growth stage-related parameters (Supplementary Fig. S3). Among the 20 stress-related libraries, biotic stress-related libraries contributed more ESTs to the contigs than libraries for abiotic stress (Supplementary Fig. S3A). Among the growth stages, libraries of the spikelet at late flowering contributed more than those of other stages, suggesting differential gene expression among the cultivars with or without any stress (Supplementary Fig. S3B). The high number of unigenes and GC% also suggests the possibility of more genes present in wheat than in other crops.[45] This is supported by the recent study of the megabase level sequencing in 3B, which

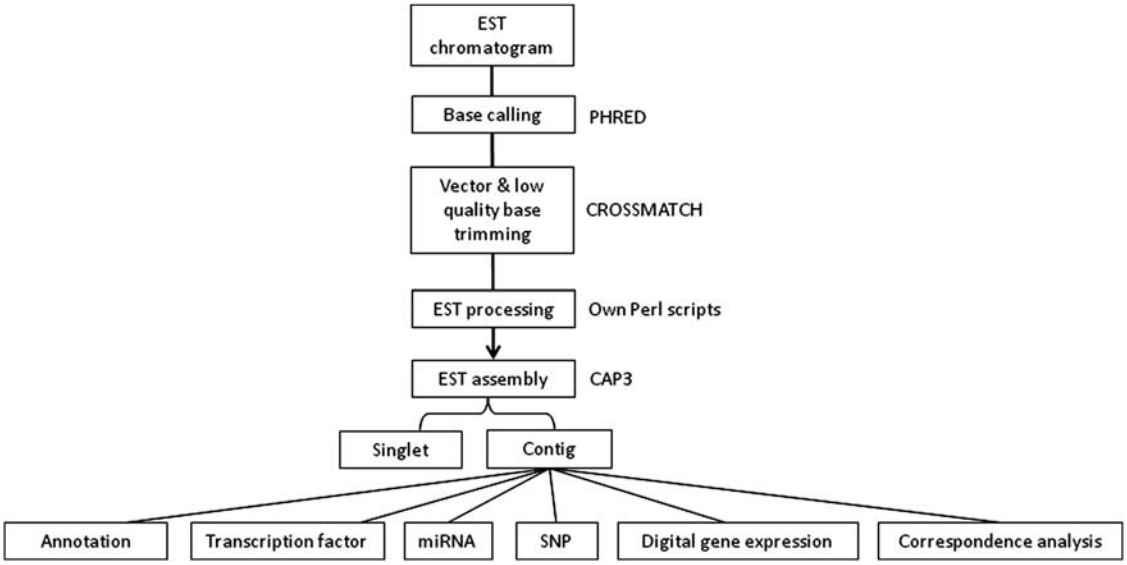

Figure 7. Schematic diagram explaining the comprehensive EST analysis. The software used in the respective step was mentioned in parallel.

estimated 50 000 genes per diploid genome as a result of the additional non-collinear genes interspersed within the highly conserved ancestral grass gene backbone, suggesting accelerated evolution in the Triticeae lineages.[44] The presence of additional genes was further confirmed by identification of new genes based on functional annotation (Fig. 3). When putative wheat gene sequences were analysed for ORF length based on their hit status, we observed significantly shorter ORFs in sequences with no hits. These results suggest that ORF length, not sequence length, is a better indicator of finding transcripts with protein coding capacity and subsequently getting a hit in a sequence database. On the other hand, more than one-third of the sequences without a hit still contained an ORF >300 bp, suggesting that sequences without a hit but with a relatively long ORF may represent new genes with protein coding capacity. We confirmed this by finding more full-length cDNA sequences. Our results also showed a higher no hit percentage in singlet sequences, most likely due to the fact that singletons represent rare genes in the wheat genome that are not well described in other organisms.

4.1. *Functional characterization*

GO analysis revealed expected categories such as molecular, biological, and cellular processes (Fig. 5). In wheat, the major molecular processes were binding and catalytic activities, similarly found in other Poaceae species.[44,46,47] Metabolic, transport, and translation functions accounted for 50% of the biological processes. Among cellular processes, as expected, membrane, intracellular, and ribosome functions played a major role. Using data from the plant TF database,[34] we found 1183 contigs from wheat that had a high similarity with 2197 different coding sequences of TF from seven species; the high similarity percentage was found with rice. The number of TF found in wheat contigs was higher than reported in *Salvia sclorea* calyx which was sequenced using 454 pyrosequncing.[48] The most represented TF family in wheat was *CCAAT*. The *CCAAT* box is a common *cis*-acting element found in the promoter and enhancer regions of a large number of genes in higher eukaryotes (for review).[49,50] In addition to this TF family, several other TF families known to be involved in plant development were also present in our data.

The role of miRNAs in developmental and stress regulation is not well established in wheat, and increasingly tissue-specific and developmental regulation of miRNAs is being found mostly in animal species.[51] Through cDNA sequencing efforts, we have identified transcripts that encode 945 different miRNAs,[52] although additional wheat-specific miRNAs may still remain in the cDNA collection. Indeed, a large number of ncRNAs have the potential to form miRNA-like stem loop precursors (data not shown), but experimental validation of these potential miRNA is required. In our study, we found an abundance of miRNA 172 target sites in as many as 236 contigs, and their uniform expression irrespective of growth stage and/or stress was confirmed by *in silico* gene expression analysis (Supplementary Fig. S2D).

The digital gene expression pattern of genes related to biotic and abiotic stresses (OMT, DREB, and NAC TFs), and epigenetic gene silencing has helped us to determine their quantitative and qualitative gene expression patterns.[7] This approach permits both the association of tissues via their common patterns of gene expression and the association of genes via their tissue-dependent expression patterns. The correlation clustering of 51 libraries formed the groups of the libraries based on respective treatments or stages which confirmed the importance of libraries as well gene expression specific to treatment.[26]

CA is an explorative computational method for the study of associations between variables. Much like principle component analysis, it displays a low dimensional projection of the data, e.g. into the plane with three-dimensional view (Supplementary Fig. S4), which can be achieved for two to three variables simultaneously, thus revealing associations between them. Traditionally, CA has been used prevalently in categorical data in the social sciences, but its application has been extended also to physical quantities and to proteomics.[53] This method allows us to quickly analyse the set of EST libraries and to discover molecular pathogenicity on wheat. In four disease-related libraries, ribulose 1−5 biphosphate and *S*-adenosyl methionine genes were commonly found. In leaf rust, UDP-glucosyl transferase and chlorophyll *a*−*b* binding protein were highly expressed,[29] while in powdery mildew treatment, ADP-ribosylation factor, lipid transfer protein, and oxalate oxidase were specifically expressed. The advantage of CA analysis helped to find the common molecular resistance mechanism in wheat.[54] While comparing susceptible and resistance reaction mechanisms, some of the genes, such as 40S, 60S ribosomal protein and Zinc finger domain-containing protein genes, have copy number variations between the two mechanisms. We have shown that the application of CA to EST data provides an informative and concise means of visualizing these data, being capable of uncovering relationships both among either gene and between genes, in particular or common stages.

4.2. SNP mining

Assembly of EST sequences into contigs in a polyploidy species like hexaploid wheat results in each contig being composed of ESTs from homoeologous loci and members of gene families.[23] SNPs are the most abundantly found co-dominant polymorphic sites in greater proportion both in intronic and in exonic regions of the genome. They occur with variable frequencies and have become very popular in plant genetics and breeding due to their amenability for high throughput genotyping. In continuation of our earlier study to discriminate homoeologous gene expression of hexaploid wheat by SNP analysis of contigs grouped from 10 cDNA libraries from Chinese Spring,[43] we have mined our large-scale data to find SNPs from 10 different cultivars with various stress treatments. With relaxed criteria, we estimated ∼50 000 SNPs in wheat with an SNP frequency of one SNP per 96 bp—a number that is higher than our previous report.[43] This high number might be due to the EST originating from 10 different cultivars with various stress conditions, although the possibilities of over-estimation from the end sequences could not be excluded. Hence, a new criterion was applied by calculating the significant sequence length to avoid the end sequence and sequencing error-based SNPs. This approach accounted for only 20% of the SNPs obtained by the initial approach, and could lead to an underestimation of nucleotide diversity, although it guarantees the elimination of false positives. The stringent parameter resulted in the SNP frequency of one SNP in every 483 bp. In comparison of SNPs among huge genome size species, coffee has one SNP every 222 bp;[55] sugarcane has one SNP every 290 bp;[16] cotton has one SNP every 500 bp;[17] oak has a frequency similar to cotton with one SNP every 471 bp.[18] Our result in wheat compared with other species could explain the low polymorphism found in polyploidy species. The comparison of SNPs found between and within cultivars showed higher SNP frequency between cultivars (Supplementary Fig. S1), confirming that the low level of polymorphism identified between homoeologous genomes compared with inter-genome differences could be useful to select parents for linkage mapping studies.

5. Conclusion

The global wheat EST assembly presented here provides an unprecedented look at the wheat transcriptome and contributes tools for wheat genetics and genomics effort. The development and inclusion of cDNA libraries from all growth stages, various tissues, and treatments portray the complete picture of wheat transcriptome. Functional annotation and characterization give new ideas about wheat expressed genome, at least in part. The identified SNPs are invaluable resources for functional genomics and molecular breeding application. This set of processed EST sequences provides a seed for future investigation of wheat functional genomics using both long and short oligonucleotide arrays. Our data will thus act as a backbone for wheat genome sequence assembly, which is progressing rapidly.

Acknowledgements: We thank Prof. K. Murai, Fukui Prefectural University, Prof. N. Kawakami, Meiji University, Dr S. Nasuda, Kyoto University, Drs S. Takumi and Y. Tosa, Kobe University, Dr T. Sutton and Prof. P. Langridge, University of Adelaide, and Dr. T. Sasaki, Okayama University, for supplying extracted RNA samples.

Supplementary Data: Supplementary Data are available at www.dnaresearch.oxfordjournals.org.

Funding

This work was supported by Grants-in-Aid for Scientific Research on priority areas 'Comparative Genomics' and the National Bio-resource Project from the Ministry of Education, Culture, Sports, Science and Technology of Japan. This is contribution No. 1007 from the Kihara Institute for Biological Research, Yokohama City University.

References

1. Braun, H.J., Atlin, G. and Payne, T. 2010, Multi location testing as a tool to identify plant response to global climate change. In: Reynolds, C.R.P. (Ed.), *Climate Change and Crop Production*. CABI: London, UK.
2. Rosegrant, M.W. and Agcaoili, M. 2010, *Global Food Demand, Supply, and Price Prospects to 2010*. International Food Policy Research Institute: Washington, DC.
3. Wilson, I.D., Barker, G.L.A., Beswick, R.W., et al. 2004, A transcriptomics resource for wheat functional genomics, *Plant Biotech.*, **2**, 495–506.
4. Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P. and Trent, J.M. 1999, Expression profiling using cDNA microarrays, *Nat. Genet.*, **21**(Suppl. 21), 10–4.
5. Schlueter, J.A., Dixon, P., Granger, C., et al. 2004, Mining EST databases to resolve evolutionary events in major crop species, *Genome*, **47**, 868–76.
6. Fulton, T.M., Hoeven, R., Eannetta, N.T. and Tanksley, S.D. 2002, Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants, *Plant Cell*, **14**, 1457–67.
7. Ewing, R.M., Kahla, A.B., Poirot, O., Lopez, F., Audic, S. and Claverie, J.M. 1999, Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression, *Genome Res.*, **9**, 950–9.
8. Ronning, C.M., Stegalkina, S.S., Ascenzi, R.A., et al. 2003, Comparative analyses of potato expressed sequence tag libraries, *Plant Physiol.*, **131**, 419–29.
9. Hughes, A. and Friedman, R. 2004, Expression patterns of duplicate genes in the developing root in *Arabidopsis thaliana*, *Mol. Evol.*, **60**, 247–56.
10. Michalek, W., Weschke, W., Pleissner, K.P. and Graner, A. 2002, EST analysis in barley defines a unigene set comprising 4,000 genes, *Theor. Appl. Genet.*, **104**, 97–103.
11. Wisman, E. and Ohlrogge, J. 2000, Arabidopsis microarray service facilities, *Plant Physiol.*, **124**, 1468–71.
12. Kawasaki, S., Borchert, C., Deyholos, M., et al. 2001, Gene expression profiles during the initial phase of salt stress in rice, *Plant Cell*, **13**, 889–905.
13. Alba, R., Fei, Z., Payton, P., et al. 2004, ESTs, cDNA microarrays, and gene expression profiling: tools for dissecting plant physiology and development, *Plant J.*, **39**, 697–714.
14. Arpat, A., Waugh, M., Sullivan, J.P., et al. 2004, Functional genomics of cell elongation in developing cotton fibers, *Plant. Mol. Biol.*, **54**, 911–29.
15. Close, T.J., Wanamaker, S.I., Caldo, R.A., et al. 2004, A new resource for cereal genomics: 22K barley GeneChip comes of age, *Plant Physiol.*, **134**, 960–8.
16. Vettore, A.L., Silva, F.R., Kemper, E.L., et al. 2003, Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane, *Genome Res.*, **13**, 2725–35.
17. Udall, J.A., Swanson, J.M., Haller, K., et al. 2006, A global assembly of cotton ESTs, *Genome Res.*, **16**, 441–50.
18. Ueno, S., Provost, G., Léger, V., et al. 2010, Bioinformatic analysis of ESTs collected by Sanger and pyrosequencing methods for a keystone forest tree species: oak, *BMC Genomics*, **11**, 650.
19. The Arabidopsis Genome Initiative. 2000, Analysis of the genome sequence of the flowering plant, *Arabidopsis thaliana, Nature*, **408**, 796–815.
20. Remington, D.L., Thornsberry, J.M., Matsuoka, Y., et al. 2001, Structure of linkage disequilibrium and phenotypic associations in the maize genome, *Proc. Natl Acad. Sci. USA*, **98**, 11479–84.
21. Tenaillon, M.I., Swakins, M.C., Long, A.D., Gaut, R.L., Doebley, J.F. and Gaut, B.S. 2001, Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* L.), *Proc. Natl Acad. Sci. USA*, **98**, 9161–6.
22. Rafalski, A. 2002, Applications of single nucleotide polymorphisms in crop genetics, *Curr. Opin. Plant Biol.*, **5**, 94–100.
23. Somers, D.J., Kirkpatrick, R., Moniwa, M. and Walsh, A. 2003, Mining single-nucleotide polymorphisms from hexaploid wheat ESTs, *Genome*, **49**, 431–7.
24. Ogihara, Y., Mochida, K., Nemoto, Y., et al. 2003, Correlated clustering and virtual display of gene expression patterns in the wheat life cycle by large-scale statistical analyses of expressed sequence tags, *Plant J.*, **33**, 1001–11.
25. Ogihara, Y., Mochida, K., Kawaura, K., et al. 2004, Construction of a full-length cDNA library from young spikelets of hexaploid wheat and its characterization by large-scale sequencing of expressed sequence tags, *Genes Genet. Syst.*, **79**, 227–32.
26. Mochida, K., Kawaura, K., Shimosaka, E., et al. 2006, Tissue expression map of a large number of expressed sequence tags and its application to *in silico* screening of stress response genes in common wheat, *Mol. Genet. Genomics*, **276**, 304–12.
27. Kawaura, K., Mochida, K., Enju, A., et al. 2009, Assessment of adaptive evolution between wheat and rice as deduced from the full-length cDNA sequence

data and the expression patterns of common wheat, *BMC Genomics*, **18**, 271.
28. Kawaura, K., Mochida, K., Yamazaki, Y. and Ogihara, Y. 2006, Transcriptome analysis of salinity stress responses in common wheat using a 22k oligo-DNA microarray, *Funct. Integr. Genomics*, **6**, 132–42.
29. Manickavelu, A., Kawaura, K., Oishi, K., et al. 2010, Comparative gene expression analysis of susceptible and resistance near-isogenic lines in common wheat infected by *Puccinia triticina*, *DNA Res.*, **17**, 211–22.
30. Ewing, B., Hillier, L., Wendl, M.C. and Green, P. 1998, Base-calling of automated sequencer traces using PHRED. I. Accuracy assessment, *Genome Res.*, **8**, 175–85.
31. Huang, X. and Madan, A. 1999, CAP3: a DNA sequence assembly program, *Genome Res.*, **9**, 868–77.
32. Gene Ontology Consortium. 2008, The Gene Ontology project in 2008, *Nucleic Acids Res.*, **36**, D440–4.
33. Quevillon, E., Silventoinen, V., Pillai, S., et al. 2005, InterProScan: protein domains identifier, *Nucleic Acids Res.*, **33**, W116–20.
34. Perez-Rodriguez, P., Riano-Pachon, D.M., Correa, L.G.G., Rensing, S.A., Kersten, B. and Mueller-Roeber, B. 2009, PlnTFDB: updated content and new features of the plant transcription factor database, *Nucleic Acids Res.*, doi: 10.1093/nar/gkp805.
35. Matsumoto, T., Tanaka, T., Sakai, H., et al. 2011, Comprehensive sequence analysis of 24,783 barley full-length cDNAs derived from 12 clone libraries, *Plant Physiol.*, **156**, 20–8.
36. Griffiths-Jones, S., Saini, H.K., Van Dongen, S. and Enright, A.J. 2008, MiRBase: tools for microRNA genomics, *Nucleic Acids Res.*, **36**, D154–8.
37. Zhang, J., Zeng, R., Chen, J., Liu, X. and Liao, Q. 2008, Identification of conserved microRNAs and their targets from *Solanum lycopersicum* Mill, *Gene*, **423**, 1–7.
38. Xie, F.L., Huang, S.Q., Guo, K., et al. 2007, Computational identification of novel microRNAs and targets in *Brassica napus*, *FEBS Lett.*, **581**, 1464–74.
39. Milne, I., Bayer, M., Cardle, L., et al. 2010, Tablet-next generation sequence assembly visualization, *Bioinformatics*, **26**, 401–2.
40. Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. 1998, Cluster analysis and display of genome-wide expression patterns, *Proc. Natl Acad. Sci. USA*, **95**, 14863–8.
41. Hamada, K., Hongo, K., Suwabe, K., et al. 2011, OryzaExpress: an integrated database of gene expression networks and omics annotations in rice, *Plant Cell Physiol.*, **52**, 220–9.
42. Choulet, F., Wicker, T., Rustenholz, C., et al. 2010, Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces, *Plant Cell*, **22**, 1686–701.
43. Mochida, K., Yamazaki, Y. and Ogihara, Y. 2003, Discrimination of homoeologous gene expression in hexaploid wheat by SNP analysis of contigs grouped from a large number of expressed sequence tags, *Mol. Genet. Genomics*, **270**, 371–7.
44. Zhang, D., Choi, D.W., Wanamaker, S., et al. 2004, Construction and evaluation of cDNA libraries for large-scale expressed sequence tag sequencing in wheat (*Triticum aestivum* L.), *Genetics*, **168**, 595–608.
45. Rabinowicz, P.D., Citek, R., Budiman, M.A., et al. 2005, Differential methylation of genes and repeats in land plants, *Genome Res.*, **15**, 1431–40.
46. Kikuchi, S., Sathoh, K., Nagata, T., et al. 2003, Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice, *Science*, **301**, 376.
47. Alexandrov, N.N., Brover, V.V., Freidin, S., et al. 2009, Insights into corn genes derived from large-scale cDNASequencing, *Plant Mol. Biol.*, **69**, 179–94.
48. Legrand, S., Valot, N., Nicolé, F., et al. 2010, One-step identification of conserved miRNAs, their argets, potential transcription factors and effector genes of complete secondary metabolism pathways after 454 pyrosequencing of calyx cDNAs from the Labiate (*Salvia sclarea* L.), *Gene*, **450**, 55–62.
49. Mitchell, P.J. and Tjian, R. 1989, Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins, *Science*, **245**, 371–8.
50. Johnson, P.F. and McKnight, S.L. 1989, Eukaryotic transcriptional regulatory proteins, *Annu. Rev. Biochem.*, **58**, 799–839.
51. Hubbard, S.J., Grafham, D.V., Beattie, K.J., Overton, I.M., McLaren, S.R. and Croning, M.D. 2005, Transcriptome analysis for the chicken based on 19,626 finished cDNA sequences and 485,337 expressed sequence tags, *Genome Res.*, **15**, 174–83.
52. Dai, X., Zhuang, Z. and Zhao, P.X. 2011, Computational analysis of miRNA targets in plants: current status and challenges, *Brief Bioinform.*, **21**, 115–21.
53. Fellenberg, K., Hauser, N.C., Brors, B., Neutzner, A., Hoheisel, J.D. and Vingron, M. 2001, Correspondence analysis applied to microarray data, *Proc. Natl Acad. Sci. USA*, **98**, 10781–6.
54. Yano, K., Imai, K., Shimizu, A. and Hanashita, T. 2006, A new method for gene discovery in large-scale microarray data, *Nucleic Acids Res.*, **34**, 1532–9.
55. Vidal, R.O., Mondego, J.M., Pot, D., et al. 2010, A high-throughput data mining of single nucleotide polymorphisms in coffea species expressed sequence tags suggests differential homeologous gene expression in the allotetraploid (*Coffea arabica*), *Plant Physiol.*, **154**, 1053–66.